

RESEARCH

Open Access



Forecasting malaria cases using climate variability in Sierra Leone

Saidu Wurie Jalloh^{1,2*}, Boniface Malenje^{3†}, Herbert Imboga^{3†} and Mary H. Hodges^{4†}

Abstract

Background Malaria continues to pose a public health challenge in Sierra Leone, where timely and accurate forecasting can guide more effective interventions. Although seasonal models such as Seasonal Autoregressive Integrated Moving Average (SARIMA) have traditionally been employed for disease forecasting, Artificial Neural Networks (ANNs) have gained attention for capturing complex temporal patterns that linear models may not fully capture.

Methods This study compares the forecasting performance of SARIMA and ANN models in forecasting malaria cases using malaria case data from 2018 to 2023. A baseline SARIMA model was developed and improved with exogenous climatic variables (precipitation, maximum temperature, and mean relative humidity) to form a SARIMAX approach. In parallel, an ANN was trained solely on historical malaria cases, without adding climatic variables.

Results SARIMA offered reasonable predictive capabilities but was outperformed by the ANN, which captured complex temporal patterns more effectively, decreasing forecast errors and improving its coefficient of determination (R^2). The SARIMA model achieved an MAPE of 12.01%, which improved further to an MAPE of 11.45% with the inclusion of climatic variables. A strong positive correlation between precipitation ($r = 0.68$) and malaria cases was observed, while maximum temperature showed a moderate negative correlation ($r = -0.45$), and mean relative humidity demonstrated a moderate positive correlation ($r = 0.55$). The ANN model outperformed both the baseline SARIMA and SARIMAX models with the lowest MAPE of 6.68%.

Conclusions These findings underscore the ANN's ability to capture non-linear dynamics, even without explicit climate inputs. These results reinforce the value of machine learning modelling approaches in guiding malaria control strategies, particularly in high-burden settings like Sierra Leone.

Keywords Malaria, Sierra Leone, Time series forecasting, Seasonal autoregressive integrated moving average, Artificial neural network, Public health, Disease surveillance

[†]Boniface Malenje, Herbert Imboga, Mary H. Hodges have contributed equally to this work.

*Correspondence:

Saidu Wurie Jalloh

wurie.saidu@students.jkuat.ac.ke

¹ Department of Mathematics (Data Science Option), Pan African University Institute for Basic Sciences Technology and Innovation, Kiambu, 00200 Juja, Kenya

² Institute for Development (IfD), New England Ville, Freetown, Sierra Leone

³ Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Kiambu, Juja, Kenya

⁴ School of Community Health, Njala University, Bo, Sierra Leone



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

Malaria remains a global health challenge, particularly in tropical and subtropical regions, where socio-economic and environmental factors compound the complexity of managing vector-borne diseases. Low-income countries, such as Sierra Leone, face a disproportionate burden of malaria, which continues to be a leading cause of morbidity and mortality, particularly among vulnerable groups such as pregnant women and children under five years of age [1]. In these regions, infectious diseases, including malaria, are major contributors to high mortality rates compared to developed nations. Due to its widespread prevalence, the disease caused by *Plasmodium* parasites and transmitted through the bites of infected *Anopheles* mosquitoes remains a pressing public health concern.

The relationship between malaria and poverty is cyclical and deeply interwoven. Malaria exacerbates poverty by reducing productivity and increasing healthcare costs, while poverty in turn limits access to essential preventive measures and treatment, fostering the spread of the disease. Despite global initiatives and efforts to reduce the malaria burden, including the Millennium Development Goals (MDGs) and the Sustainable Development Goals (SDGs), the persistence of the disease underscores the evident socio-economic disparities in its incidence. Individuals in urban areas with access to better healthcare are at a lower risk compared to those in impoverished rural settings, where resources for prevention and treatment are scarce. This highlights the need for comprehensive public health interventions that target socio-economic determinants such as income inequality, living conditions, and healthcare accessibility.

In Sierra Leone, malaria control faces numerous challenges, including weak health infrastructure, environmental variability, and economic constraints. The country's entire population is at risk, with malaria historically classified as hyper-endemic, showing a prevalence of over 50% [2]. As of 2021, malaria prevalence had decreased to 22% among children aged 6–59 months [3], and the malaria incidence was reported at 308.14 cases per 1000 people at risk in 2022 [1]. The dominant malaria parasite is *Plasmodium falciparum*, responsible for over 90% of cases, and the main mosquito vectors, *Anopheles gambiae* and *Anopheles funestus*, have developed resistance to several insecticides, further complicating control efforts [2].

Sierra Leone's health infrastructure has been severely impacted by years of civil war, the Ebola epidemic, and the COVID-19 pandemic, leaving a fragile healthcare system with limited access to services. The three-tier healthcare delivery system, composed of primary, secondary, and tertiary care, is anchored by 1288 peripheral health units and supported by 8159 community health workers

[2]. These workers play an essential role in maternal and child health services, particularly after the implementation of the revised Community Health Workers strategy in 2022 [2]. However, despite initiatives like the Free Health Care Initiative launched in 2010, which aimed to provide free treatment to pregnant women, lactating mothers, and children under five, access to care remains inconsistent, hampered by hidden costs and logistical challenges.

There has been notable progress in malaria control efforts, as indicated by survey data showing an increase in the use of insecticide-treated nets (ITNs) and a reduction in malaria prevalence. From 2016 to 2021, the proportion of households owning at least one ITN rose marginally from 60% to 61%, while the percentage of children under five sleeping under ITNs increased significantly from 44% to 76%. Concurrently, the prevalence of parasitaemia among children under five decreased from 40% in 2016 to 22% in 2021 [2]. These trends suggest some success in reducing malaria transmission rates, but the disease burden remains high, requiring sustained and intensified control measures.

Environmental factors play a crucial role in malaria transmission, influencing both mosquito breeding and human vulnerability to infection. Climate change, deforestation, and poor waste management practices have created environments conducive to mosquito proliferation, intensifying malaria transmission in many regions. Studies have shown strong correlations between malaria incidence and environmental factors such as temperature, precipitation, vegetation cover, and proximity to water bodies [4, 5]. For instance, the Normalized Difference Vegetation Index (NDVI) has been identified as a reliable predictor of malaria risk, with higher NDVI values often correlating with increased transmission [6]. Poor waste management in urban areas exacerbates the situation, creating mosquito breeding grounds and raising malaria incidence, as evidenced by research in Nigeria and Burkina Faso [7, 8]. Understanding these environmental factors is critical for developing effective malaria prevention and control strategies, particularly in regions with seasonal transmission surges.

Climatic variability influences malaria transmission, as fluctuations in temperature, precipitation, and humidity affect mosquito breeding and parasite development. For instance, moderate temperature increases generally enhance malaria transmission by accelerating mosquito and parasite life cycles, whereas excessively high temperatures reduce mosquito survival, thus decreasing transmission [9]. Additionally, humidity levels above 60% have been shown to promote malaria transmission [9, 10]. Rainfall patterns also affect mosquito breeding, with malaria transmission peaking at the onset and end of the

rainy season, typically in May and October–November in Sierra Leone. The interaction between these environmental factors and human behaviour further complicates malaria control efforts.

Predictive modelling plays an essential role in managing malaria by enabling timely responses based on anticipated transmission patterns. Seasonal Autoregressive Integrated Moving Average (SARIMA) models have been successfully applied to forecast malaria cases in various settings. These models have demonstrated reasonable accuracy in various regions, including the United States, South Africa, Iran, and Kenya, with studies reporting Mean Absolute Percentage Errors (MAPE) as low as 21.6% [11–14]. SARIMA models are particularly useful for tailoring interventions to local conditions, as different regions require specific parameter adjustments to account for local climate and transmission dynamics [12, 14].

In recent years, advancements in Machine Learning (ML) and Artificial Intelligence (AI) have further enhanced malaria prediction capabilities. Artificial Neural Networks (ANNs), in particular, excel at capturing complex, non-linear relationships in time-series data, making them highly effective for forecasting malaria incidence. These models learn from historical data and have been shown to outperform traditional statistical methods like logistic regression in several studies [15]. For instance, a study in Somaliland found that ANNs provided superior predictive accuracy when meteorological variables were included [16]. Furthermore, the integration of genetic algorithms with ANNs has demonstrated even greater accuracy in evolving prediction models [17].

ANNs also hold promise for diagnostic applications in malaria-endemic regions with limited access to laboratory facilities. By analysing patient history, symptoms, and environmental data, ANN-based models can offer rapid, reliable predictions of malaria cases. Some models, when combined with image processing techniques like Speeded-Up Robust Features (SURF), have achieved high accuracy in detecting malaria parasites in blood smear images, with sensitivity and specificity rates exceeding 85% [18]. Mobile applications leveraging these AI-driven tools could offer community health workers in remote areas an accessible means of diagnosis and case management, significantly contributing to malaria control efforts.

This study builds on these developments, utilising both SARIMA and ANN models to forecast malaria cases in Sierra Leone based on historical incidence. Comparison of the performance of these models aims to identify the most effective approach for forecasting malaria cases, thereby supporting the development of targeted and timely public health interventions.

Methods

Study area

Sierra Leone is a West African nation bordering Liberia to the south and Guinea to the north and east [19, 20]. It spans about 72,000 km² with a diverse landscape ranging from sea level up to over 1600 m, as shown in Fig. 1 [21]. Along the Atlantic coast, the coastal plains extend roughly 320 km and comprise low-lying wetlands and mangrove swamps that typically remain below 40 km in width [19]. The highest elevations are found in the eastern and northeastern Interior Plateau [20].

Average monthly temperatures are relatively stable around the year, ranging from 25 °C to 35 °C. The dry season sees higher temperatures, while the wet season is slightly cooler. The harmattan (December–February) brings cooler conditions with dry, dusty winds [22]. Humidity fluctuates seasonally from about 70% in the dry season to around 80% in the wet season, and drops during the harmattan [23].

Data collection

Secondary data from January 2018 to December 2023 was used, primarily from Sierra Leone's Health Management Information System (HMIS), which includes laboratory-confirmed malaria cases from health facilities nationwide. Meteorological data, including mean relative humidity, precipitation, wind speed, solar radiation, and maximum temperature, were collected from the Sierra Leone Meteorological Agency. However, due to missing data points, the dataset was supplemented with corresponding data obtained from the World Weather Online (WWO) website for the period 2018 to 2023. The WWO data was carefully processed and integrated to ensure a comprehensive and accurate dataset for subsequent analysis.

While this supplementation provided the necessary data to perform the analysis, it is important to acknowledge potential limitations arising from integrating data from multiple sources. WWO often relies on models, satellite data, or interpolation techniques to estimate weather variables, particularly for regions with limited ground-based observations. As a result, variability in data collection methods between WWO and the Sierra Leone Meteorological Agency may introduce inconsistencies that could affect the precision of the models. To mitigate these risks, the supplemented data were cross-validated with available local sources where possible.

Elevation data were sourced from the Global Solar Atlas Terrain Elevation layer (May 2024), derived from SRTM v4.1, viewfinderpanoramas.org, and the GEBCO 2014 Grid. This dataset offers a 30 arc-second spatial resolution and global geographic coverage from –60° to 65° latitude

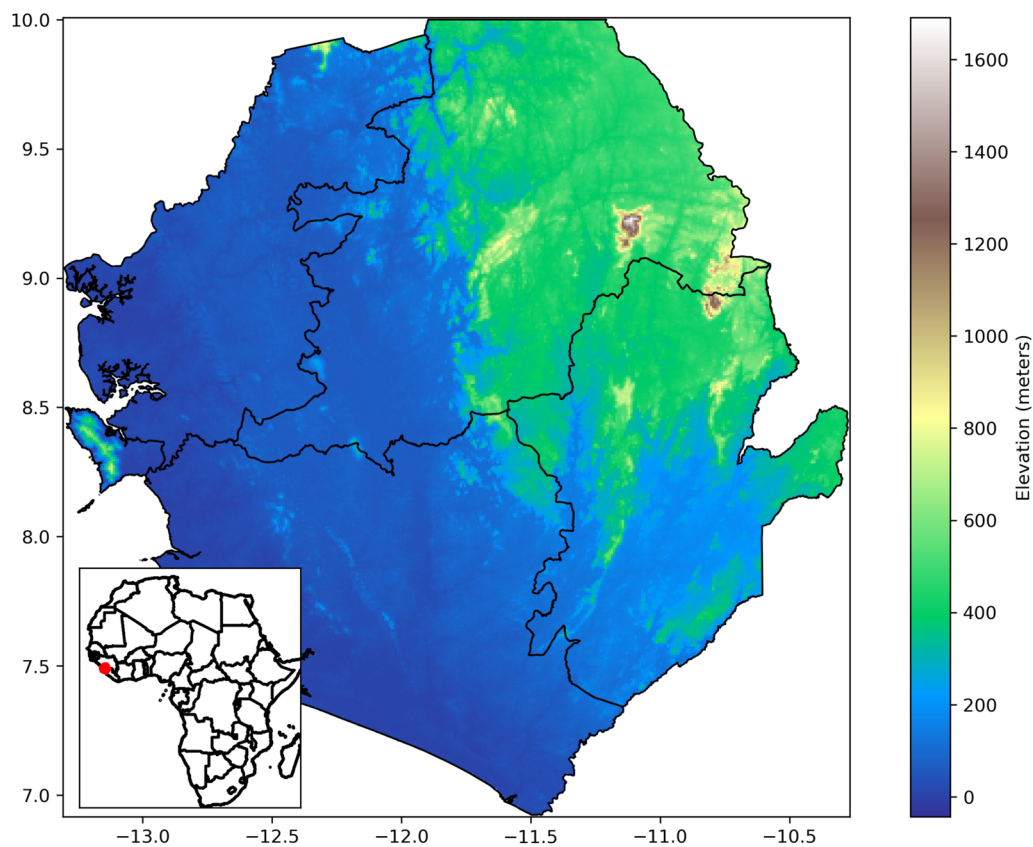


Fig. 1 Elevation Map of Sierra Leone [21]

and -180° to 180° longitude, licensed under Creative Commons (CC BY 4.0).

Data from the 2021 Sierra Leone Malaria Indicator Survey were also used to provide context on malaria prevalence and intervention coverage. A total of 72 monthly data points were analysed for each variable in the analysis.

Data processing

Precipitation, maximum temperature, mean relative humidity, and malaria case data operate on different scales and exhibit varying distributions; so, an unequal normalisation strategy was used to preserve these distinctions and reduce the risk of any single variable dominating the model. For the malaria case data, a log transformation was employed:

$$\text{Log}(\text{Malaria Cases}) = \log(\text{Malaria Cases} + 1),$$

which stabilises variance, mitigates the impact of extreme spikes, and ensures that zero-valued observations remain valid after transformation. In contrast, precipitation and mean relative humidity were min-max scaled:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}},$$

where $X_{\min}$ and $X_{\max}$ denote the minimum and maximum values of each feature in the training subset. These variables often exhibit skewed distributions and benefit from being mapped to a fixed [0, 1] range. The maximum temperature was standardised:

$$X_{\text{standard}} = \frac{X - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation in the training subset. This choice better accounts for its relatively narrower, more symmetric distribution. This variable-specific normalisation helped capture the unique patterns inherent to malaria incidence and each climate feature, thereby strengthening the subsequent modelling process.

Outlier detection

Outliers in the dataset can distort predictions and lead to inaccurate models. To mitigate these effects, a two-step outlier detection strategy was implemented that combines both univariate and multivariate techniques.

An initial exploration of the data using box plots and scatter plots was conducted to visually inspect for anomalies. Each numeric feature was analysed independently using the $\mu \pm 2\sigma$ rule. Any observation falling outside the range defined was flagged as a univariate outlier. This step quickly identifies extreme values in individual features that deviate from the expected range.

In parallel, a clustering-based approach was used to capture multivariate outliers, which are observations that may not appear extreme when each variable is considered separately but are unusual in the context of the combined feature space. The procedure began by standardising the relevant variables to ensure equal weighting in subsequent distance calculations. As an initial guideline, the number of clusters, k was estimated using the formula:

$$k = \sqrt{\frac{n}{2}},$$

where $n = 72$ (the total number of observations), yielding an estimate of $k = 6$. This provided a reference for further empirical validation. However, the final selection of k was based on data-driven techniques.

For each data point x_i in the dataset (where $i = 1, 2, \dots, 72$), the K-means algorithm calculates the Euclidean distance between the data point and each of the k cluster centroids. The Euclidean distance between a data point x_i and the centroid C_j of cluster j is given by:

$$d(x_i, C_j) = \sqrt{(x_{i1} - C_{j1})^2 + (x_{i2} - C_{j2})^2 + \dots + (x_{im} - C_{jm})^2},$$

where $x_{i1}, x_{i2}, \dots, x_{im}$ are the values of the features for the data point x_i , $C_{j1}, C_{j2}, \dots, C_{jm}$ are the corresponding values for the centroid C_j , and m is the number of features in the dataset.

Once all data points are assigned to their nearest clusters, each centroid is updated as the mean of all points in its respective cluster:

$$C_j(i) = \frac{1}{|S_j|} \sum_{x \in S_j} x_i,$$

where S_j is the set of data points assigned to cluster j , $|S_j|$ is the number of data points in cluster j , and x_i is the value of the i -th feature for each point in S_j .

After clustering, the distance from each data point to its assigned cluster centroid is calculated. Within each cluster, points whose distance exceeded the 95th percentile of the distance distribution were flagged as outliers. This multivariate method captures observations that, although not flagged by univariate criteria, deviate from the typical feature profile of their cluster.

Finally, the two approaches were combined by considering a data point as an outlier if it was flagged by either

the univariate or the clustering-based method. This strategy ensures that both obvious univariate anomalies and more subtle multivariate outliers are effectively identified.

Collinearity test

To assess collinearity, the Variance Inflation Factor (VIF) was used, which quantifies how much the variance of a regression coefficient is inflated due to the correlation of a predictor with the other predictors. A VIF greater than 5 was considered problematic because it indicates a high level of multicollinearity, which can lead to unreliable statistical inferences.

Pearson correlation

The Pearson correlation between each climatic variable and the malaria case data was calculated. Variables with a correlation coefficient greater than 0.6 or less than -0.6 were deemed to have a strong relationship with malaria transmission.

Seasonal autoregressive integrated moving average

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the traditional ARIMA framework to account for seasonality, which is a common feature in time series data such as monthly malaria cases. It is mathematically expressed as

$$SARIMA(p, d, q) \times (P, D, Q)_s.$$

The SARIMA model is represented using the backshift operator B , which shifts the time series back by one period:

$$B^k y_t = y_{t-k}.$$

Using this notation, the general SARIMA model is written as

$$\Phi_P(B^s) \phi_p(B) (1 - B)^d (1 - B^s)^D y_t = \Theta_Q(B^s) \theta_q(B) \epsilon_t.$$

In this formulation, the functions $\phi_p(B)$ and $\theta_q(B)$ represent the non-seasonal autoregressive (AR) and moving average (MA) components, respectively, and are defined as

$$\begin{aligned} \phi_p(B) &= 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p, \\ \theta_q(B) &= 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q. \end{aligned}$$

Likewise, the functions $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ represent the seasonal AR and seasonal MA components, respectively, and are given by

$$\begin{aligned}\Phi_P(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps}, \\ \Theta_Q(B^s) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs}.\end{aligned}$$

The operator $(1 - B)^d$ is the non-seasonal differencing operator that ensures stationarity in the non-seasonal domain, while $(1 - B^s)^D$ is the seasonal differencing operator that ensures stationarity in the seasonal domain. The term ϵ_t represents a white noise with mean zero and constant variance.

Log transformation and lagged features

In time series forecasting, large spikes in data can disproportionately influence model predictions, leading to suboptimal parameter estimation. Mathematically, the transformation is defined as:

$$\text{Log}_y = \log(y_t + 1),$$

where Log_y represents the log-transformed malaria cases, y_t is the observed number of malaria cases at time t . Adding 1 prevents numerical issues when $y_t = 0$, ensuring stability in the transformation. While log transformation alone does not guarantee stationarity, it reduces variance instability, making subsequent differencing more effective when necessary.

An autocorrelation analysis on historical malaria case data was conducted to identify significant temporal dependencies and seasonality. Strong autocorrelations were observed at one-month and 12-month lags, indicating that both short-term dependencies and seasonal transmission cycles influence malaria cases. Based on these findings, lagged values X_{t-1} , and X_{t-12} were included as exogenous regressors in the SARIMA model to capture both immediate and seasonal effects.

In addition to lagged features, a 3-month moving average was introduced to smooth fluctuations and mitigate the impact of short-term noise in the data. The selection of lag and moving average structures was guided by out-of-sample forecasting accuracy, validation metrics, and insights from the existing literature on malaria transmission cycles. Lagged values help the model learn dependencies over time, while moving averages ensure that short-term fluctuations do not distort long-term trends.

Model selection

To identify the best SARIMA configuration, an Auto-ARIMA approach was employed, which is an automated method that iteratively evaluates multiple ARIMA models and selects the best-fitting one based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The final model was

chosen based on the lowest AIC and BIC values, balancing model complexity and accuracy.

Artificial neural network

Artificial Neural Networks (ANNs) are machine learning models inspired by the human brain, effective for modelling complex, nonlinear data relationships, especially in time series forecasting. A Multilayer Perceptron (MLP) ANN was employed to forecast malaria cases using engineered temporal features and historical case data.

Feature engineering

To model temporal dependencies, lag features and moving averages were created. Lag features were created for malaria case counts from 1 to 12 months before each observation. The 12-month lag window was chosen to capture annual seasonality, a well-documented pattern in malaria transmission linked to climatic cycles and mosquito breeding seasons. These lagged values are represented as:

$$X_t = X_{t-k},$$

where X_t is the malaria case count at time t , and k is the lag period. Shorter lags capture recent transmission dynamics, while longer lags account for annual cyclical trends.

Moving averages were computed over 3-month and 6-month windows to smooth short-term noise and highlight underlying trends. The 3-month average reflects quarterly fluctuations, while the 6-month average aligns with semi-annual climatic shifts. These are defined as:

$$MA_k = \frac{1}{k} \sum_{i=1}^k X_{t-i},$$

where MA_k is the moving average over k months.

Temporal features such as month, quarter, and year were also extracted from the timestamp. Months and quarters explicitly encode seasonal patterns, while the year accounts for long-term trends unrelated to seasonality, like population growth or intervention programs.

Model optimization

The ANN was trained using the Adam optimiser, which adaptively adjusts learning rates during training. Mean Squared Error (MSE) was used as the loss function:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

where y_i is the true value and $\hat{y}_i$ is the predicted value. To prevent overfitting, L2 regularisation penalised large weights, and dropout layers randomly deactivated neurons during training. Hyperparameters (layer sizes, activation functions) were tuned using a grid search cross-validation to validate the model's ability to generalise to unseen data.

Evaluation of forecasting accuracy measure

The forecasting error for a time series $y_1, \dots, y_n$, where $\hat{y}_i$ represents the forecasted value at time i , is defined as follows:

$$e_i = y_i - \hat{y}_i, \quad i \leq n.$$

Model accuracy is evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2).

Results

Descriptive analysis

The time series plot highlights cyclical patterns in malaria cases, with notable peaks in 2019 and 2021 and dips in 2020 and 2022, as shown in Fig. 2. The mean values show an upward trend from 2018 to 2019, followed by fluctuations in subsequent years. The distribution of cases varies, with right-skewed distributions in 2019 and 2021 and left-skewed distributions in 2022 and 2023, indicating shifting conditions.

Figure 3 shows a bar chart of seasonal variation in malaria cases, with the highest cases observed in June (9.62%) and July (9.36%), coinciding with the onset of the rainy season in Sierra Leone. May, August, and September also show relatively high percentages, highlighting the rainy season as a critical period for malaria transmission due to favourable mosquito breeding conditions. Conversely, the lowest cases occur in February (6.89%), during the dry season, when mosquito populations are reduced. The months of January, March, and April also have lower transmission rates. October, November, and December show a gradual decline in cases from the mid-year peak, reflecting a lag in transmission after the rainy season.

Results of feature engineering

The relationship between climate variables and malaria cases was explored using Pearson correlation and p-value tests to identify statistically significant features. In addition, the Variance Inflation Factor (VIF) was calculated for each variable to assess multicollinearity. Table 1 summarises these results.

Precipitation showed a strong positive correlation ($r = 0.68$), indicating its significant association with higher malaria transmission due to increased mosquito breeding sites. Temperature had a moderate negative correlation ($r = -0.45$), suggesting that lower temperatures might support higher malaria cases, as extreme heat can hinder mosquito survival. Humidity exhibited a moderate positive correlation ($r = 0.55$), implying that higher humidity levels may also favour mosquito survival and thereby contribute to increased malaria incidence.

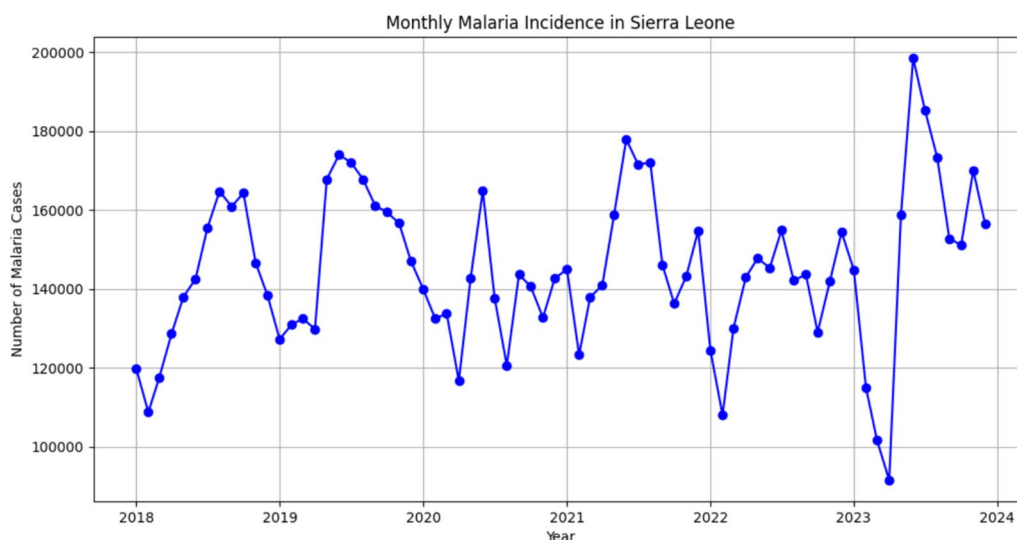


Fig. 2 Time series plot of malaria incidence

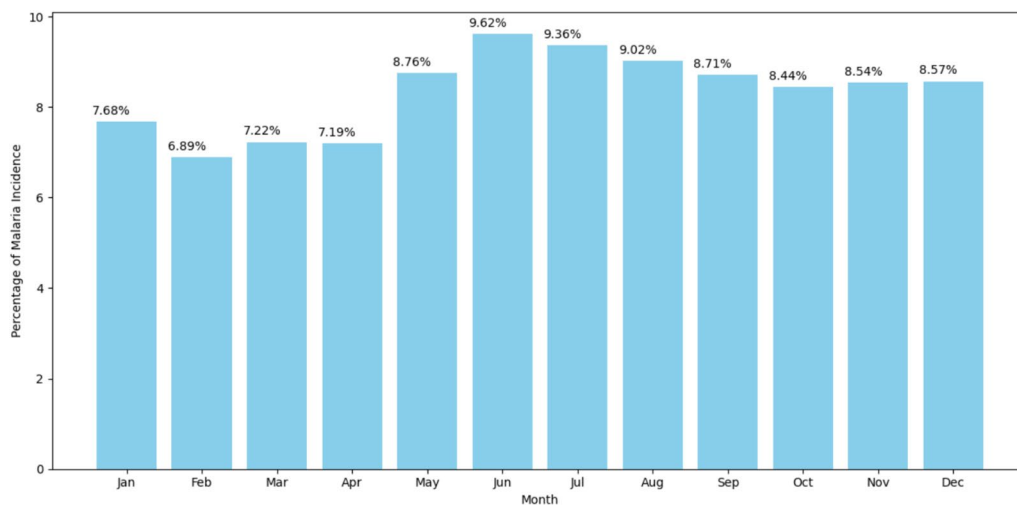


Fig. 3 Monthly percentage of malaria incidence in Sierra Leone

Table 1 Statistical significance and multicollinearity testing results

Weather variable	Pearson correlation (r)	p-value	Variance inflation factor (VIF)
Precipitation	0.68	0.003	2.5
Maximum Temperature	−0.45	0.015	3.2
Mean Relative Humidity	0.55	0.008	4.1
Wind Speed	0.12	0.320	6.8
Solar Radiation	−0.09	0.380	5.5

Wind speed ($p = 0.320$, $VIF = 6.8$) and solar radiation ($p = 0.380$, $VIF = 5.5$) had p -values above 0.05, indicating they were not statistically significant predictors. In addition, their high VIF values suggested considerable collinearity with other features. Consequently, these variables were removed in subsequent model iterations to address potential issues with inflated variance.

Outlier detection

This process identified 4 outliers in malaria cases (5.6%). Precipitation data showed 3 outliers (4.2%), all occurring in August of different years (2019, 2022, and 2023) with extremely high rainfall values exceeding 1000 mm. Maximum temperature showed 3 outliers (4.2%), all being unusually low temperatures in August of different years (2019, 2021, and 2022), consistently falling just below the lower threshold. No outliers were detected in the mean relative humidity measurements, suggesting this variable has more consistent patterns throughout the dataset.

Threshold analysis

Annual malaria incidence data for Sierra Leone (1990–2022) were obtained from the Institute for Health Metrics and Evaluation (IHME) Global Burden of Disease dataset, processed by Our World in Data [24]. To identify significant spikes, the WHO-recommended threshold of $\bar{x} + 2\sigma$ [25] was applied, where $\bar{x}$ is the mean incidence and σ is the standard deviation. The mean incidence ($\bar{x}$) and standard deviation (SD) were calculated as follows:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2},$$

where x_i is the malaria incidence in year i , $\bar{x}$ is the mean incidence, and n is the total number of observations ($n = 32$). In this analysis, the sample mean incidence was calculated to be 420.59 cases per 1000 population, with a sample standard deviation of 44.69.

Between 1990 and 2022, malaria incidence remained below the epidemic threshold of 509.96 cases per 1,000 population, as shown in Fig. 4. This suggests there were no unusual spikes in transmission, and observed annual variations stayed within expected limits.

Log transformation

Figure 5 presents the log-transformed data. The log transform visibly reduces the wide variation seen in raw counts, making seasonal trends and overall patterns more apparent. Although peaks and troughs remain visible, the overall variance is more stable.

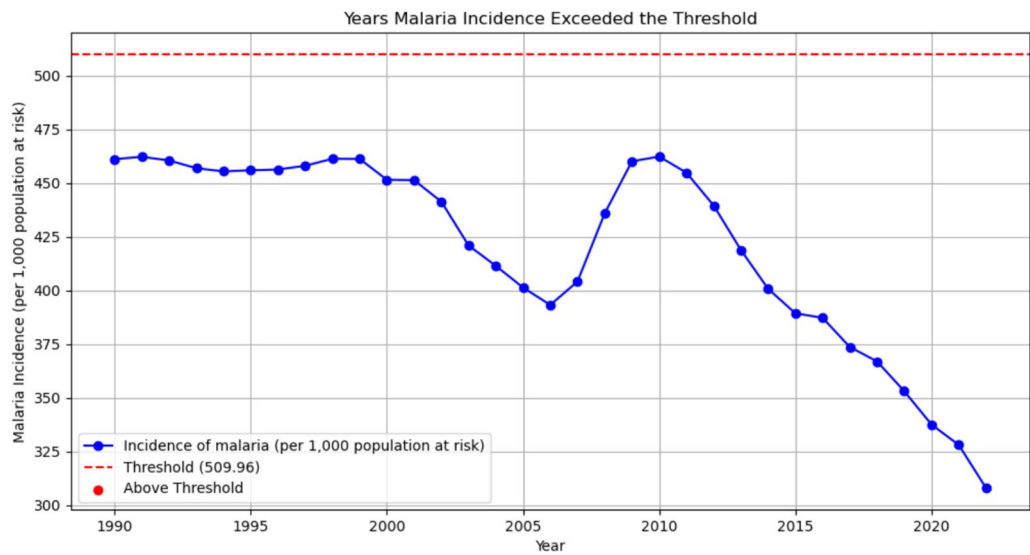


Fig. 4 Annual malaria incidence (per 1000 population)

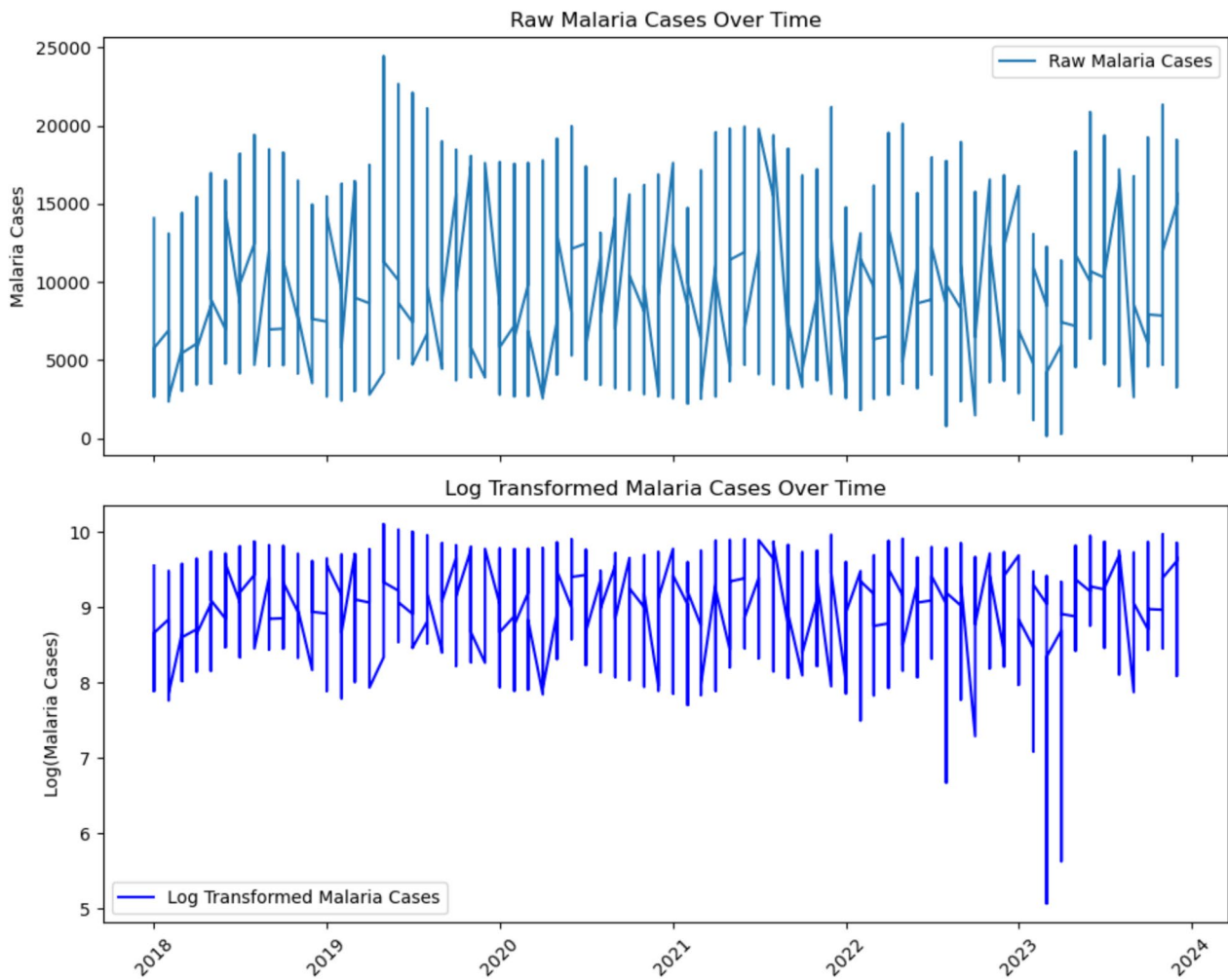


Fig. 5 Log-transformed malaria cases

Table 2 Performance comparison of SARIMA and ANN models

Model	RMSE	MAE	MAPE (%)	R^2
SARIMA model	0.132	0.111	12.01	0.59
ANN model	0.013	0.010	6.68	0.92

Table 3 Performance metrics of SARIMA with climatic variables

Model	RMSE	MAE	MAPE (%)	R^2
SARIMAX model	0.124	0.102	11.45	0.65
ANN model	0.013	0.010	6.68	0.92

Performance comparison

The SARIMA model was initially fitted to the malaria data, followed by the ANN model. Both models were evaluated to determine their effectiveness in capturing the underlying patterns in the data. The ANN model achieved the lowest errors across all metrics, higher R^2 , underscoring its superior predictive performance as shown in Table 2.

Inclusion of climatic variables

Building on the initial SARIMA model, it was observed that the ANN model outperformed SARIMA without including climate variables. To investigate whether a statistical time-series approach can achieve similar performance improvements, exogenous climatic variables were included. Precipitation, maximum temperature, and mean relative humidity were chosen based on their correlation with malaria cases, and adding these climate regressors accounts for external drivers beyond internal

seasonality. This procedure yielded a SARIMAX model that outperformed the baseline SARIMA, as indicated by lower forecast errors and improved information criteria. Although the ANN remained the top performer overall, the SARIMAX with exogenous inputs demonstrated that adding climatic variables improves forecasts.

Table 3 shows that adding precipitation, maximum temperature, and mean relative humidity to the SARIMA-based model reduced its forecast errors and raised its R^2 to 0.65. However, the ANN still performed better across all measures, with notably lower errors and a higher R^2 of 0.92.

Residuals diagnostics

A residual analysis was conducted to ensure that the model's assumptions were met. The residuals, which represent the differences between the observed values and the model predictions, should ideally be uncorrelated and normally distributed around zero. These diagnostics confirm that the residuals are reasonably normally distributed, as shown in Figs. 6, 7 and 8.

Discussion

This study performs a comparative analysis of SARIMA and Artificial Neural Network (ANN) models for forecasting malaria cases in Sierra Leone. The findings highlight the superior performance of the ANN model in forecasting malaria trends, aligning with recent studies that underscore the growing role of machine learning in disease prediction. For instance, ANNs have been shown to outperform traditional statistical models in disease forecasting, offering significant improvements in accuracy, particularly in non-linear data relationships

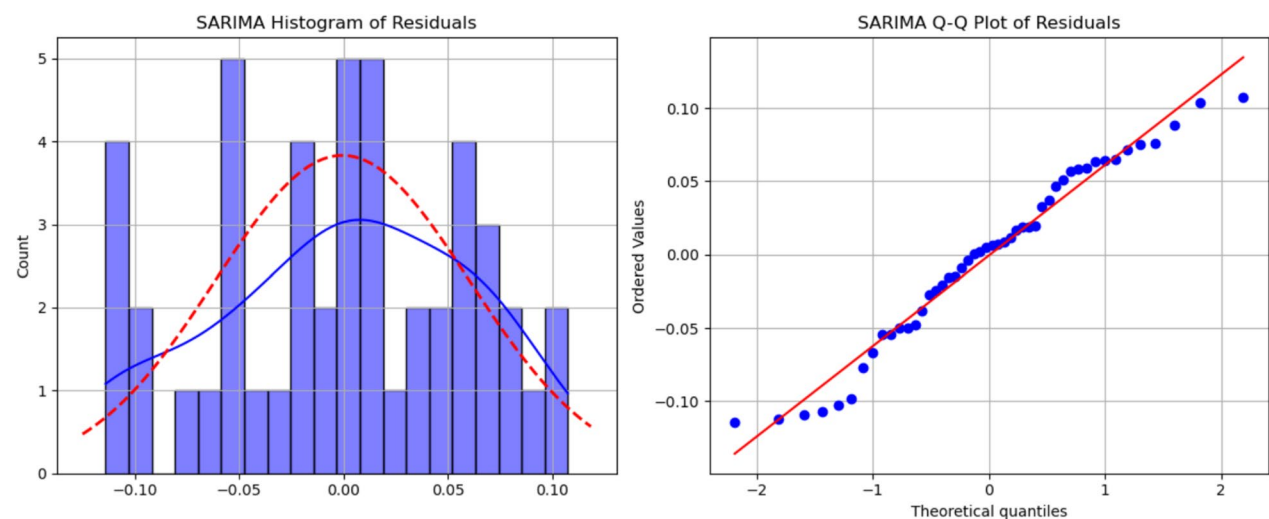


Fig. 6 Residuals of SARIMA model

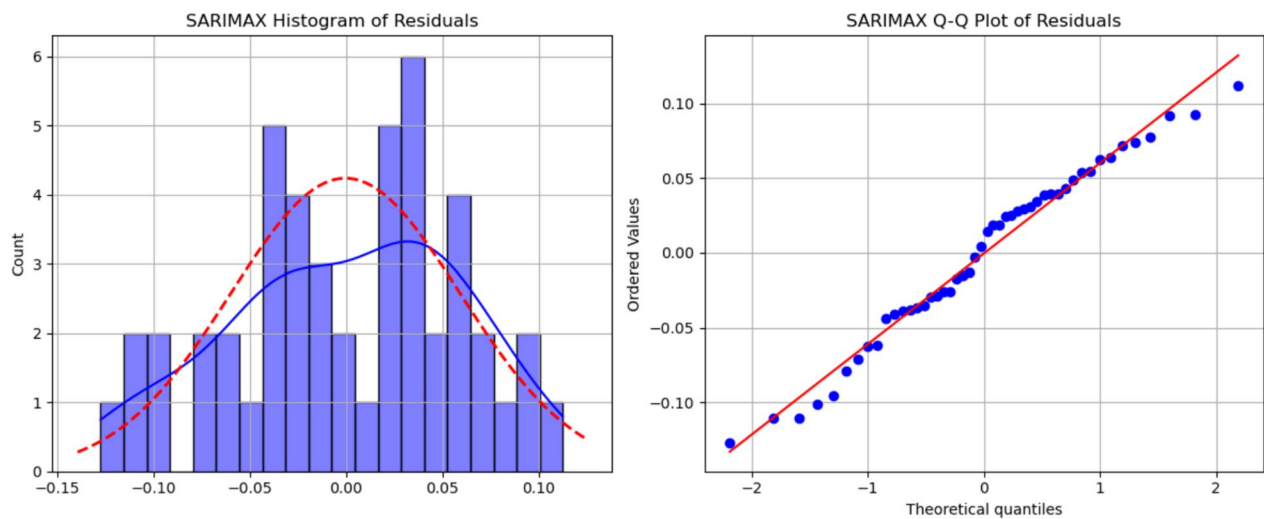


Fig. 7 Residuals of SARIMAX model

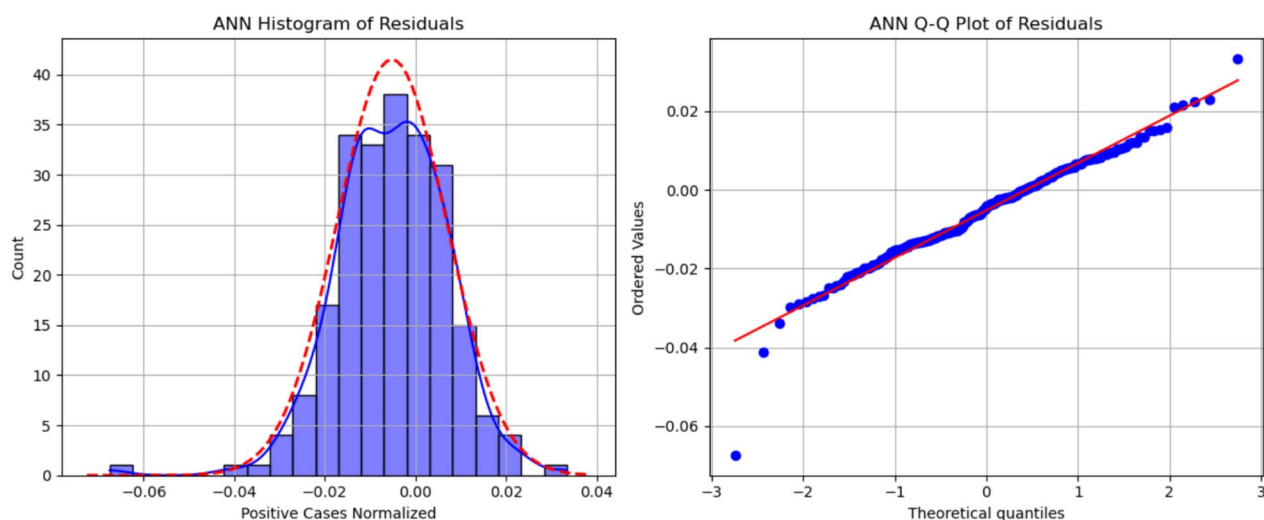


Fig. 8 Residuals of ANN model

[21, 26]. In the analysis, the ANN model outperformed the SARIMA model, which recorded an R^2 of 0.59, which increased to 0.65 when precipitation, maximum temperature, and mean relative humidity were added. However, the ANN achieved an R^2 of 0.92 even without the inclusion of climatic variables, underscoring its ability to capture non-linear data relationships and complex temporal dependencies. These results align with other studies highlighting the comparative advantage of machine learning techniques in disease forecasting [16, 21, 26] and underscore the potential value of future investigations that integrate climate variables into the ANN to further refine its performance. Recent trends indicate an

increasing burden of malaria in many regions, with a projected rise in *Plasmodium knowlesi* infections in Malaysia [27] and an increase in malaria incidence in Zambia [28]. These findings suggest that without timely interventions, malaria cases could continue to rise, driven by environmental changes and inadequate control measures.

The SARIMA model, though effective in capturing seasonal patterns, did not perform as well as the ANN model in this study. Similar applications of SARIMA for disease forecasting have been reported in several studies [11, 16, 29, 30], but machine learning models like ANNs often achieve better accuracy due to their ability to model complex, non-linear relationships. The SARIMA

model's limitations in capturing the full range of transmission dynamics highlight the need for more advanced approaches in high-burden settings. Machine learning has emerged as a powerful tool across healthcare for predictive analytics. Various supervised ML techniques, such as random forests, support vector machines (SVM), and logistic regression, have been successfully applied to predict diseases such as diabetes, liver disorders, and cancer [31]. Similar applications in malaria forecasting are likely to gain traction, especially with the integration of patient data, environmental factors, and clinical history. This study's ANN model demonstrates the promise of ML techniques in public health forecasting by achieving significant improvements in predictive accuracy.

Despite the strengths of this study, such as the inclusion of climatic variables, limitations remain. Relying on health facility data can introduce reporting biases [28, 30]. Future research should prioritise enhancing data quality and developing localised models that consider spatial heterogeneity, as malaria transmission can vary substantially within small geographic areas [16]. Although this study did not incorporate climatic variables into the ANN model, future studies should explore their integration, given the documented influence of environmental factors on malaria transmission and the anticipated impacts of climate change in regions such as sub-Saharan Africa [21]. Additionally, the model's generalizability could be limited since malaria transmission patterns differ across regions, influenced by distinct ecological and socio-economic contexts [26]. Emphasising model interpretability and rigorous validation processes will be essential for advancing machine learning applications in disease forecasting. This study demonstrates that ANN models can improve malaria forecasting accuracy in Sierra Leone, which is vital for timely interventions and aligns with ongoing global efforts toward malaria elimination.

Conclusions

In light of the study findings, the ANN model demonstrated superior predictive accuracy for forecasting malaria cases in Sierra Leone compared to traditional models like SARIMA, even after including climatic variables, achieving lower forecasting errors and higher predictive accuracy. These results highlight ANN's potential as a powerful tool for public health decision-making by enabling timely and targeted interventions. Effective malaria forecasting is crucial for Sierra Leone, where timely and precise interventions can reduce the burden of malaria. Coordinated efforts with international organisations, such as the WHO, should focus on refining predictive models into practical public health strategies through improved data collection and validation

methods, along with developing localised forecasting at the sub-national level. Strengthening current interventions, such as insecticide-treated bed net distribution and indoor residual spraying, will support Sierra Leone's long-term goal of malaria elimination.

Author contributions

Saidu Wurie Jalloh conceptualized the study, developed the methodology, and performed the formal analysis. Saidu Wurie Jalloh also managed data curation and visualization. Boniface Malenje, Herbert Imboga, and Mary H. Hodges provided supervision and guidance throughout the project.

Funding

This research was not supported by any external funding.

Data availability

The data used in this study are available upon reasonable request from the corresponding author.

Code availability

The code used for data analysis in this study is available from the corresponding author upon reasonable request.

Materials availability

The materials used in this study are available upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate

Ethical approval was not required for this study, as it utilized publicly available data with no human subjects involved.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Received: 13 September 2024 Accepted: 26 April 2025

Published online: 20 May 2025

References

1. WHO. World Malaria Report 2020. World Health Organization; 2020. Available from: <https://www.who.int/publications/i/item/9789240015791>.
2. U S President's Malaria Initiative.: Sierra Leone FY 2024 Country Malaria Profile - Final. Last updated: 05/04/2023. PDF Document. Available from: <https://www.pmi.gov/where-we-work/sierra-leone/>.
3. Leone] NMCPNS, Leone SS, of Sierra Leone U, Services CR, International U. Sierra Leone Malaria Indicator Survey 2021. Freetown, Sierra Leone: NMCP, SSL, CRS, and Utica International; 2021.
4. Okunlola OA, Oyeyemi OT. Spatio-temporal analysis of association between incidence of malaria and environmental predictors of malaria transmission in Nigeria. *Sci Rep*. 2019;9:17500.
5. Okiring J, Routledge I, Esptein A, Namuganga JF, Kanya EV, Odei Obeng-Amoako G, et al. Associations between environmental covariates and malaria incidence in high transmission settings of Uganda: a distributed non-linear lagged ecological analysis. *BMC Public Health*. 2021;21:1962.
6. Gaudart J, Touré O, Dessay N, Dicko AL, Ranque S, Forest L, et al. Modelling malaria incidence with environmental dependency in a locality of Sudanese savannah area, Mali. *Malar J*. 2009;8:61.
7. Oyeboode OJ. Solid waste management for sustainable development and public health: a case study of Lagos State in Nigeria. *Universal J Public Health*. 2013;1:33–9.

8. Kafondo P, Segda BG, Nzihou JF, Koulidiati J. Environmental impacts of waste management deficiencies and health issues: a Case Study in the City of Kaya, Burkina Faso. *J Environ Protect*. 2013;2013:37557.
9. Mohammadkhani M, Khanjani N, Bakhtiari B, Tabatabai SM, Sheikhzadeh K. The relation between climatic factors and malaria incidence in Sistan and Baluchestan, Iran. *Sage Open*. 2019;9:2158244019864205.
10. Mohammadkhani M, Khanjani N, Bakhtiari B, Sheikhzadeh K. The relation between climatic factors and malaria incidence in Kerman, South East of Iran. *Parasite Epidemiol Control*. 2016;1:205–10.
11. Permasari AE, Hidayah I, Bustoni IA. SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence. In: 2013 international conference on information technology and electrical engineering (ICITEE). IEEE; 2013. p. 203–207.
12. Ebhuoma O, Gebreslasie M, Magubane L. A seasonal autoregressive integrated moving average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa. *S Afr Med J*. 2018;108:573–8.
13. Tohidinik HR, Keshavarz H, Mohebbi M, Sanjar M, Hassanpour G. Prediction of malaria cases in the southeastern Iran using climatic variables: an 18-year SARIMA time series analysis. *Asian Pac J Trop Med*. 2021;14:463–70.
14. Berengu PT. Modelling of malaria admission cases for children under five years using Sarima model [PhD Thesis]. Meru University; 2019.
15. da Cunha GB, Luitgards-Moura JF, Naves ELM, Andrade AO, Pereira AA, Milagre ST. Utilização de uma rede neural artificial para predição da incidência de malária na cidade de Canta, estado de Roraima. *Rev Soc Bras Med Trop*. 2010;43:567–70.
16. Mohamed J, Mohamed AI, Daud EI. Evaluation of prediction models for the malaria incidence in Marodijeh Region, Somaliland. *J Parasit Dis*. 2022;46:395–408.
17. Rismala R, Liong TH, Ardiyanti A. Prediction of malaria incidence in Banggai regency using evolving neural network. In: 2013 International Conference on Technology, Informatics, Management, Engineering and Environment. IEEE; 2013. p. 89–94.
18. Özbilge E, Güler E, Güvenir M, Şanlıdağ T, Özbilgin A, Süer K. Automated malaria parasite detection using artificial neural network. In: International Conference on Theory and Applications of Fuzzy Systems and Soft Computing. Springer; 2020. p. 631–640.
19. Mansaray AS, Barrie A, Jabati EB, Juanah MS, Ganda KJ, et al. Utilizing geo-spatial tools to assign unique identifiers to water bodies in a low-income country. *J Water Resour Protect*. 2021;13:236.
20. Nicol DSWH, Fyfe C, Sesay SM. Sierra Leone. *Encyclopedia Britannica*.
21. Jalloh SW, Imboga H, Hodges MH, Malenje B. Comparative evaluation of predictive models for malaria cases in Sierra Leone. *Open J Epidemiol*. 2025;15:188–216.
22. Weather Atlas: Weather Atlas: Climate Data for 2024. Accessed 27 Aug 2024. <https://www.weather-atlas.com>.
23. World Bank: World Bank Country and Lending Groups—Country Classifications. <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>. The World Bank Group.
24. Institute for Health Metrics and Evaluation (IHME): Incidence of malaria [dataset]. Original data from IHME Global Burden of Disease - Prevalence and Incidence. IHME, Global Burden of Disease (2024)—with major processing by Our World in Data. Retrieved March 15, 2025 from <https://ourworldindata.org/grapher/the-incidence-of-malaria-per-1000-population>.
25. Roser M, Ritchie H, Ortiz-Ospina E, Hasell J. Our world in data. 2020. Burden of disease URL: <https://ourworldindata.org/burden-of-disease>[Accessed 2020-02-09][WebCite Cache]. 2020.
26. Taylor M, Bennett C, Schoen M, Hoque S. Advances in artificial neural networks as a disease prediction tool. *J Cancer Res Ther*. 2021;9:1–11.
27. Ooi CH, Bujang MA, Bakar TMITA, Ngui R, Lim YAL, et al. Over two decades of Plasmodium Knowlesi infections in Sarawak: trend and forecast. *Acta Trop*. 2017;176:83–90.
28. Menda DM, Nawa M, Zimba RK, Mulikita CM, Mwandia J, Mwaba H, et al. Forecasting confirmed Malaria cases in northwestern province of Zambia: a time series analysis using 2014–2020 routine data. *Adv Public Health*. 2021;2021: 652252.
29. Muthukumar T, Karnaboopathy R, Annadurai K, Raja T. Forecasting of malaria in India using time series analysis. *J Comprehen Health*. 2019;7:16–22.
30. Anwar MY, Lewnard JA, Parikh S, Pitzer VE. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malar J*. 2016;15:566.
31. Kumar A, Bansal H, Jaiswal A, Gupta SK. Early disease prediction using ML. *Int J Adv Eng Nano Technol*. 2023;11:123–34.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.